# Reaching Petabyte-Scale System with ZNS SSDs

2022. 11. 01 | Wonchul Lee,
Samsung Electronics

SAMSUNG

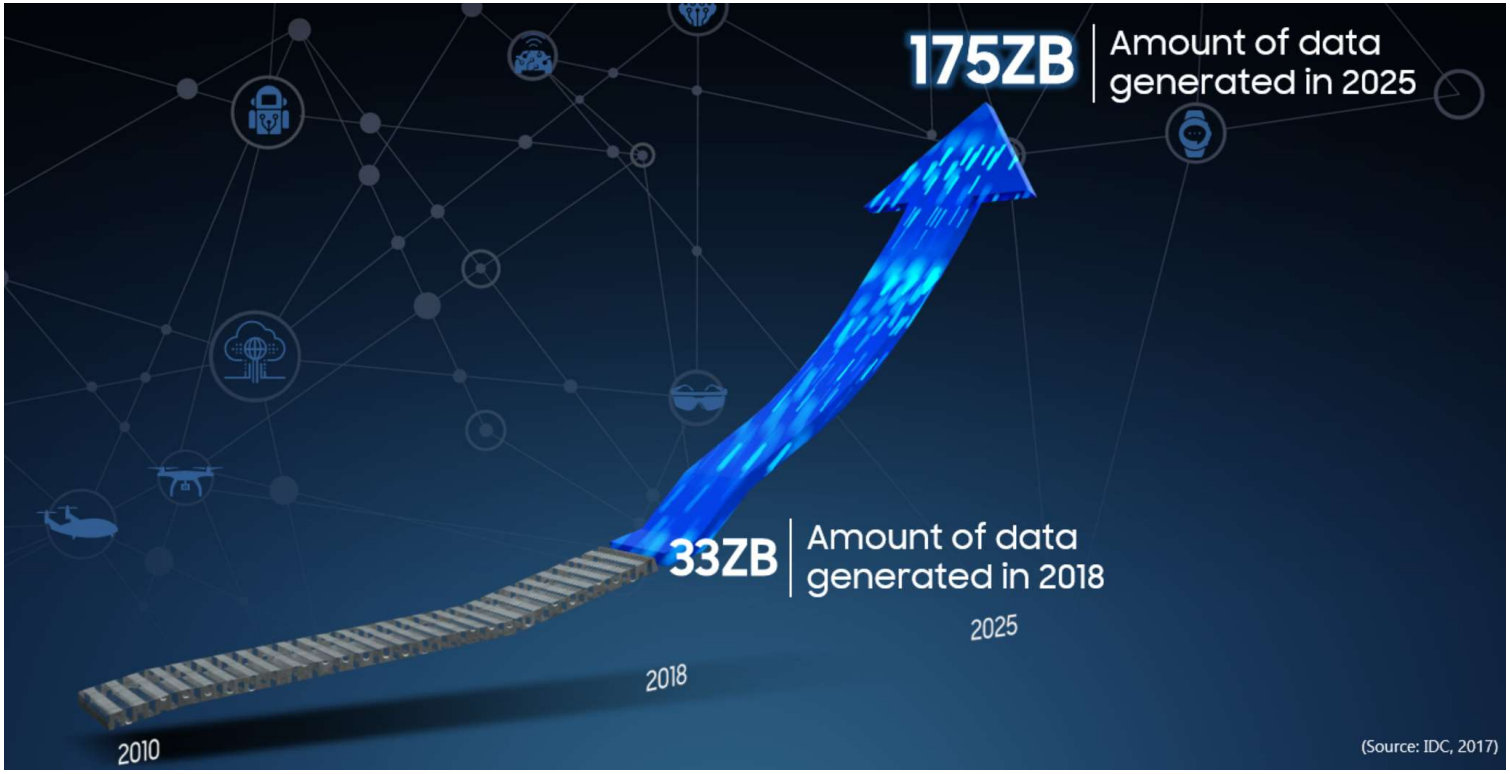
SOLUTION
Core Values

# Contents

# The Advent of the Data Age



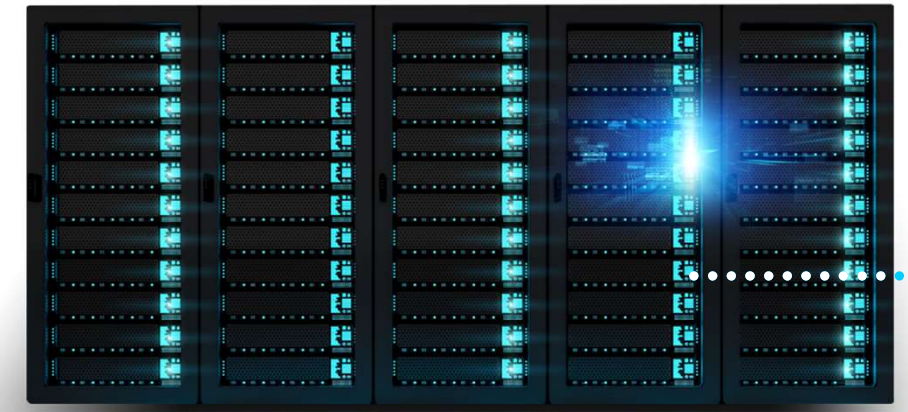
- **Data is increasing at an exponential rate and could reach 175ZB by 2025.**



# Petabyte SSD (PBSSD)



- Disaggregated NVMe subsystem with QLC SSDs increases rack-scale space and power efficiency
- Building block for a super high-density storage system

Exabyte scale storage system

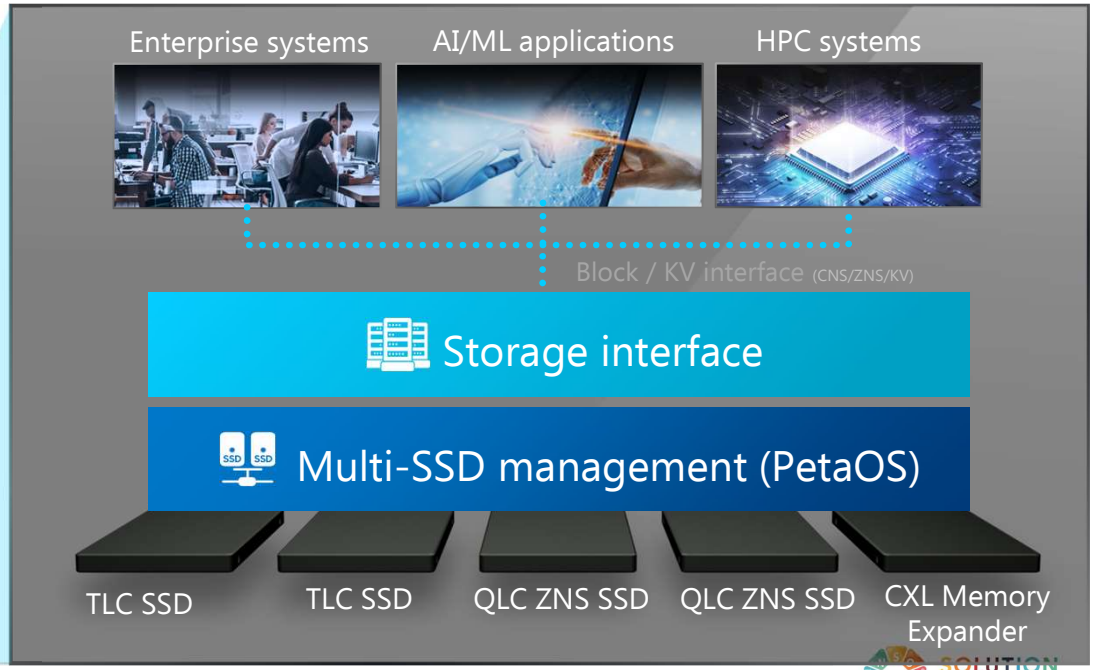
## 1PB disaggregated storage box

Space efficiency I Power efficiency I High performance

# Petabyte SSD (PBSSD)



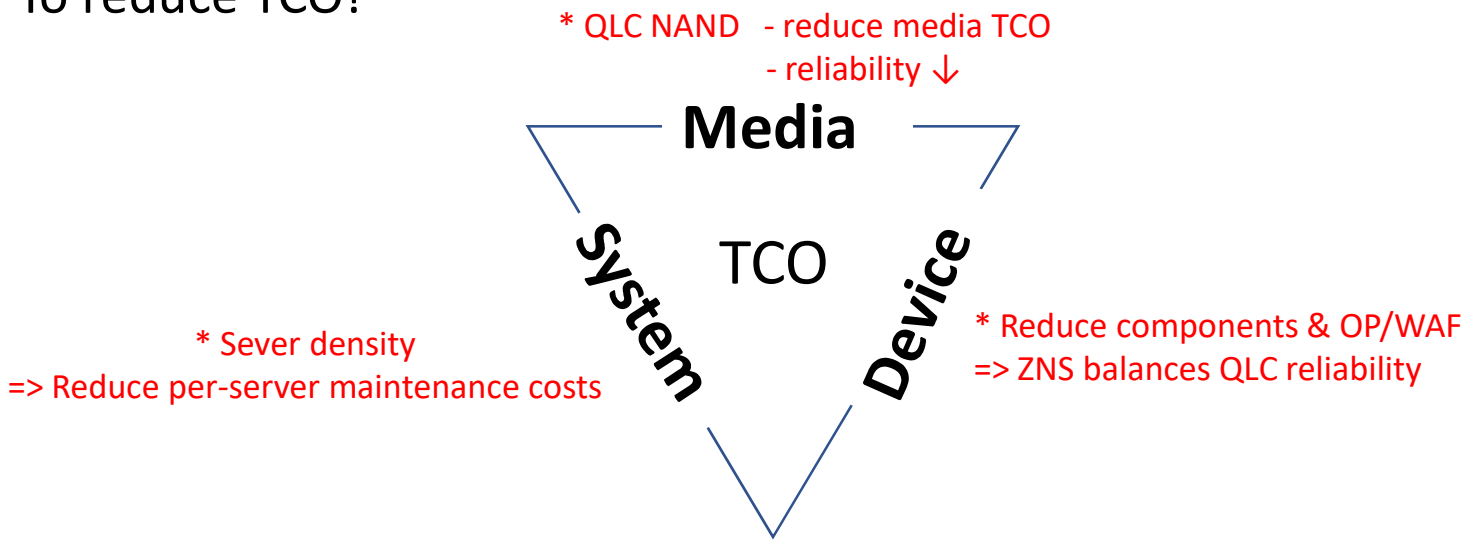
- Disaggregated NVMe subsystem with QLC SSDs increases rack-scale space and power efficiency
- Building block for a super high-density storage system



# Architecture



- To reduce TCO!



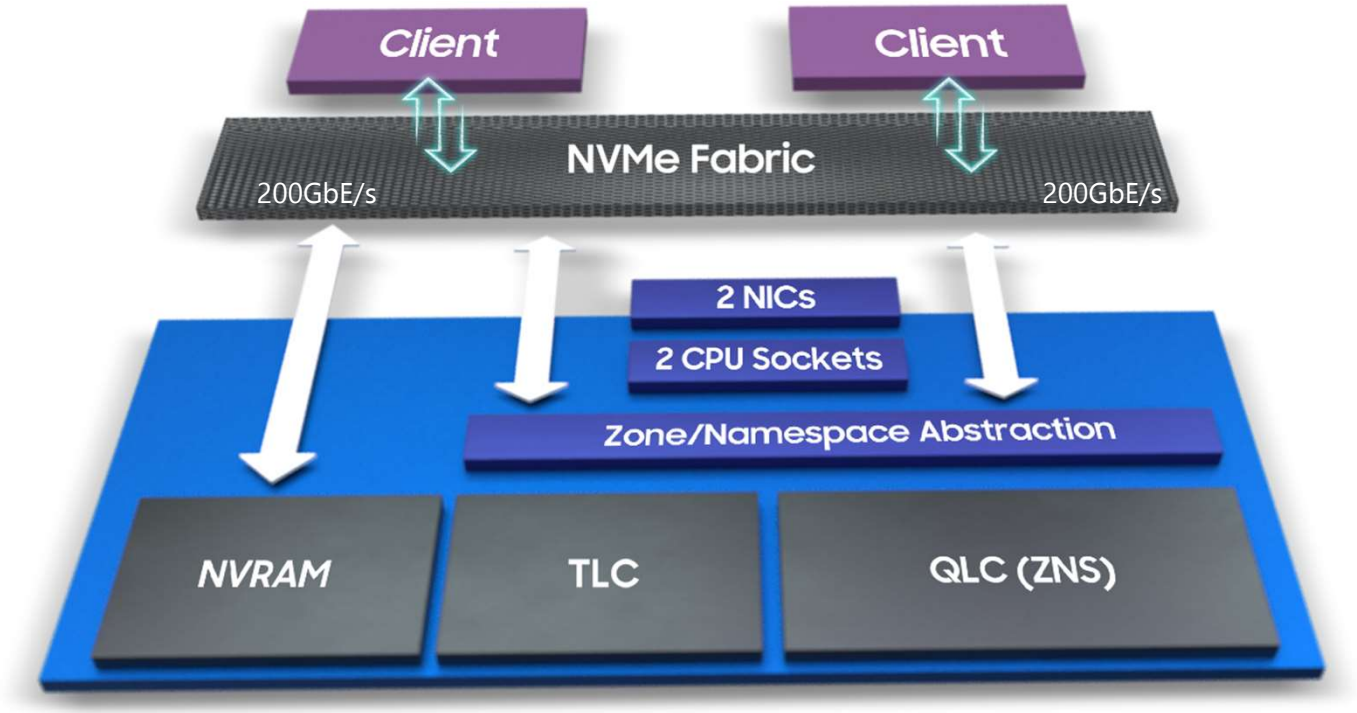

- Challenges
  - Zone & Data management O/H
  - Blast radius
  - Bring-up time

# PBSSD Reference System



- Announced at Flash Memory Summit 2022

* Poseidon V2 is announced at OCP Global Summit 2021

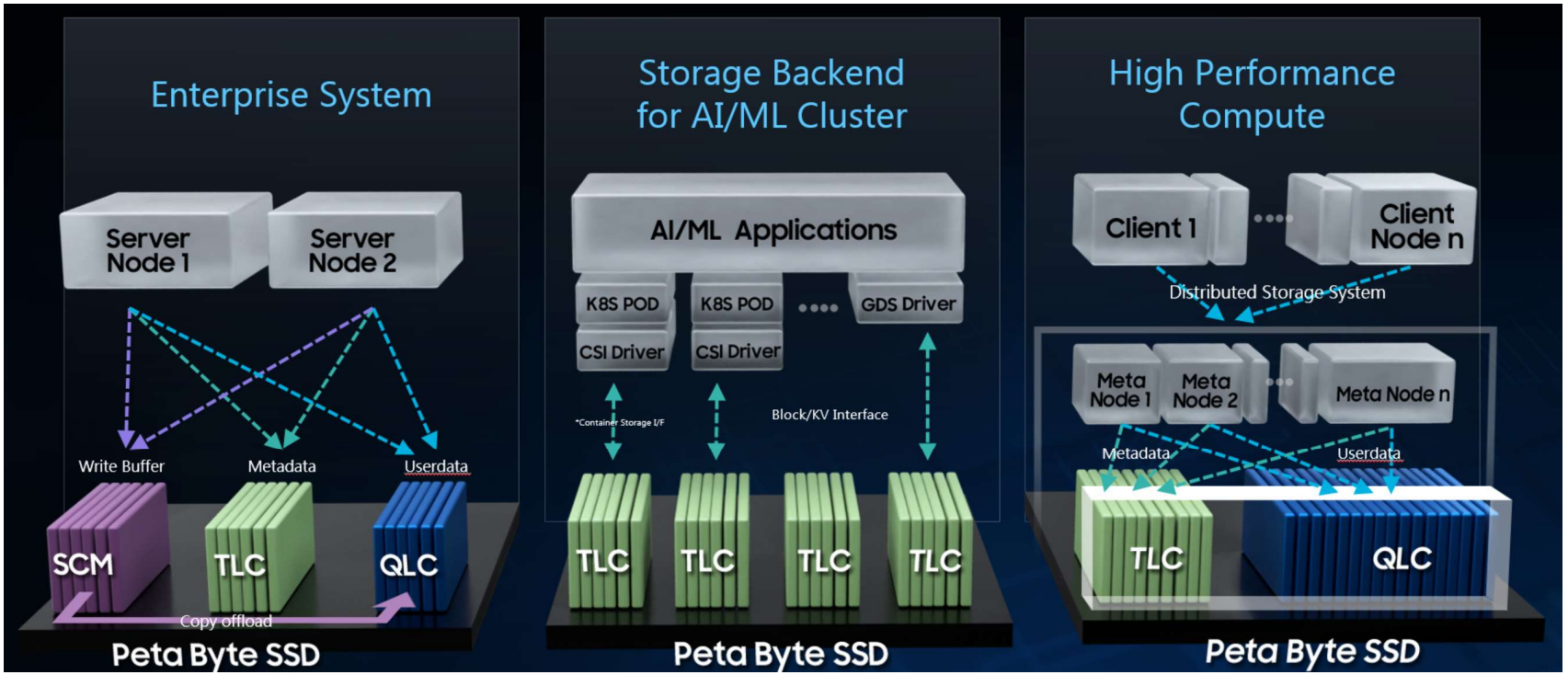



NIC : Dual 200GbE
TLC : PM1743 PCIe Gen5 V6 TLC 16TB x 8ea **(128TB)**
QLC : BM1731a V5 QLC 128TB x 16ea **(2PB)**

# Use Cases



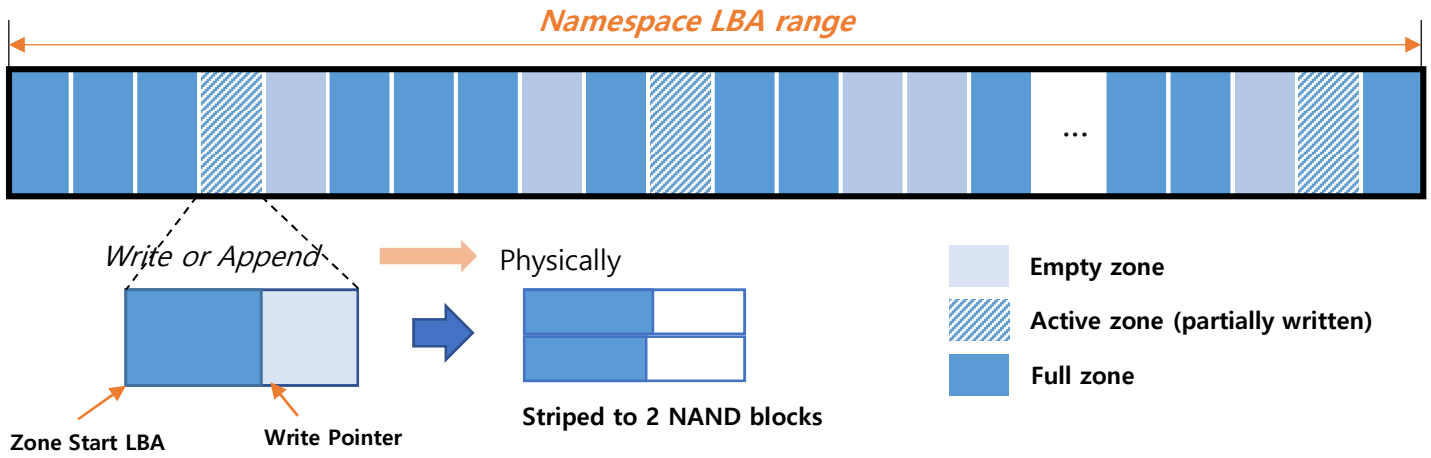
- Various models can be configured according to the SSD and S/W used.



# ZNS Concept (Overview)



- **Zoned Namespace**
  - ✓ **SSD storage space is divided into a fixed size called ZONE ( Zone =? NAND Block Size )**
  - ✓ **Each zone allows only Sequential Write**

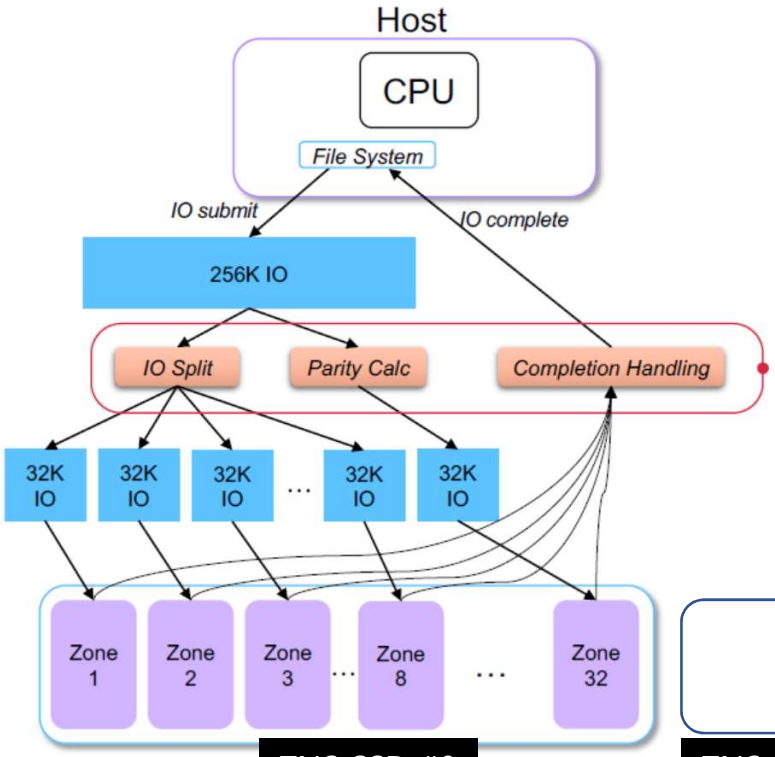



  - ✓ **Define zone information through zone descriptor and communicate with host**

# Challenge 1: Zone & Data Management



- To use ZNS SSDs, host S/W needs CPU cost to manage zones.

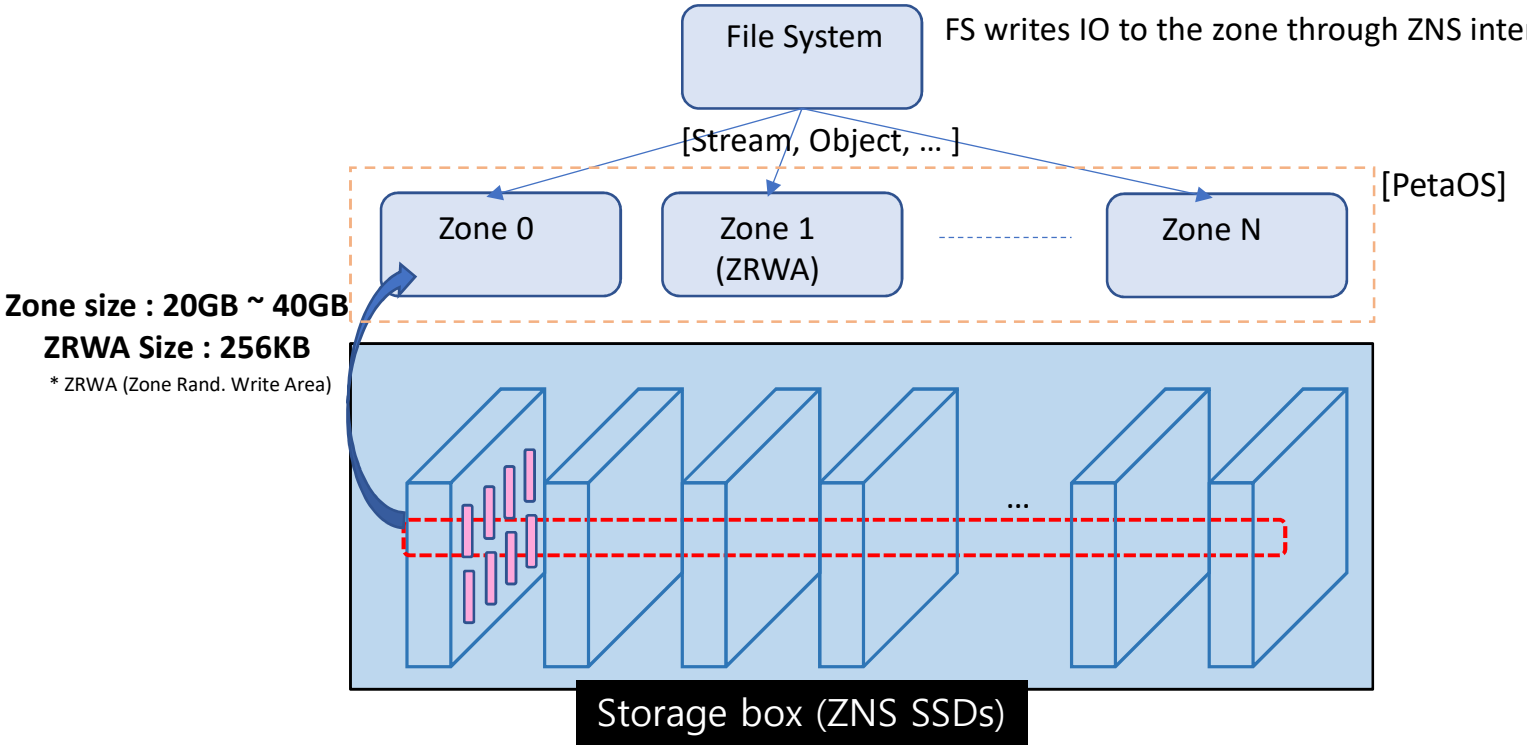





**[Host S/W]**

1. Zone management
   - Write pointer, zone activities, …
2. IO scheduling
3. EC/Parity calculations.
4. Garbage collection
5. …

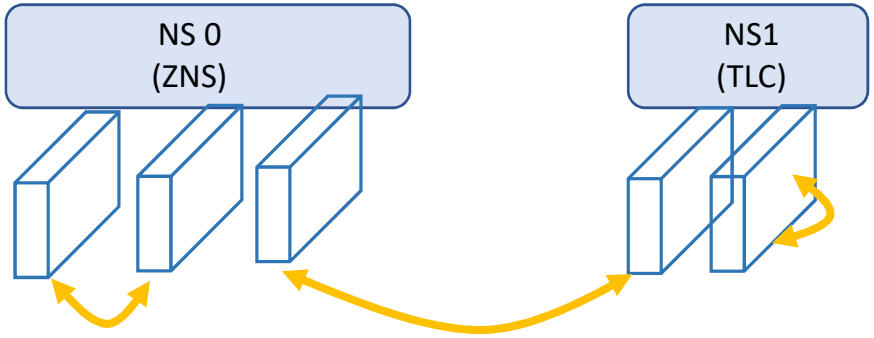
- **By offloading device (ZONE) management, it is possible to focus on the original function of host software.**



# Challenge 1 : Zone & Data Management

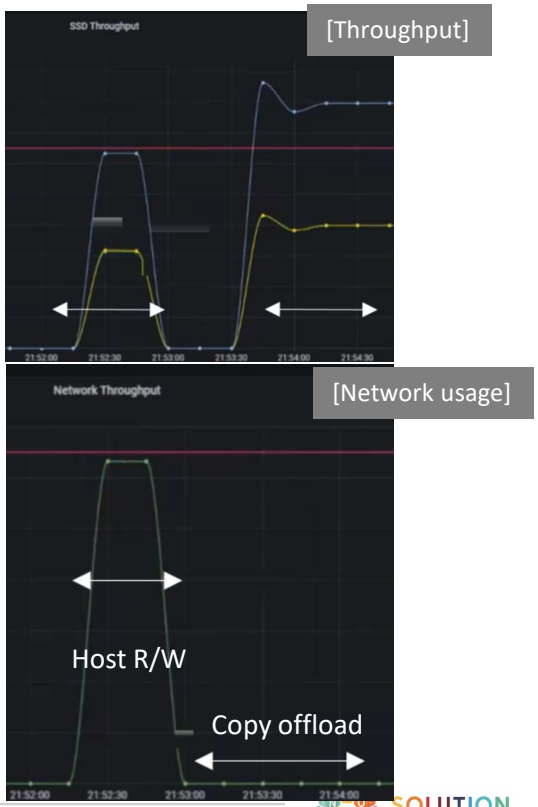


- PetaOS supports data movement between namespaces
  - Intra/Inter Copy : **Performance can exceed network max**
    - **Intra copy : data move inside the NS**
    - **Inter copy : data move between the NSs**




**[Use case]**

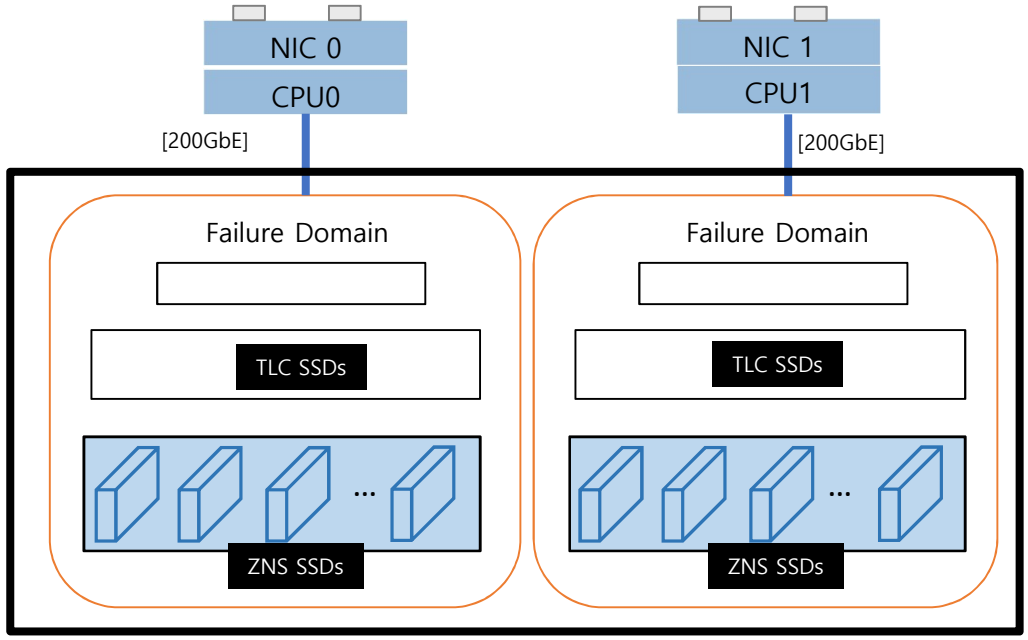
- **Control tiering with different type's storages.**
- **Replication, GC, …**

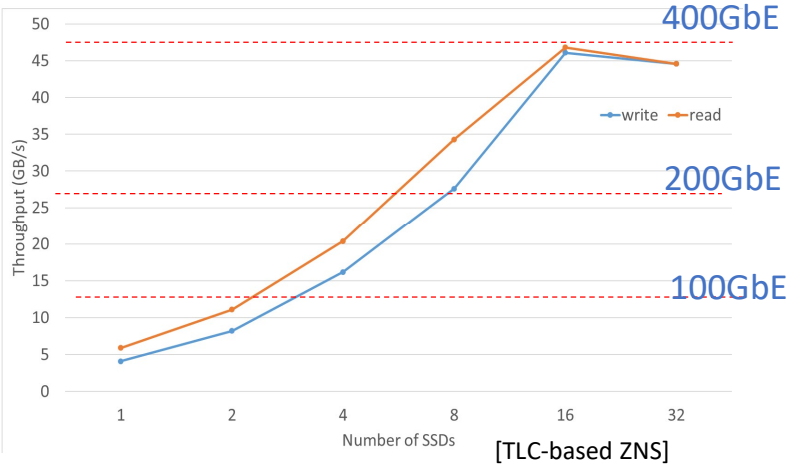


# Challenge 2 : Blast Radius



- **Failure Domain (FD) : Isolate to minimize blast radius due to component failure.**
  - How many SSDs are in a FD?



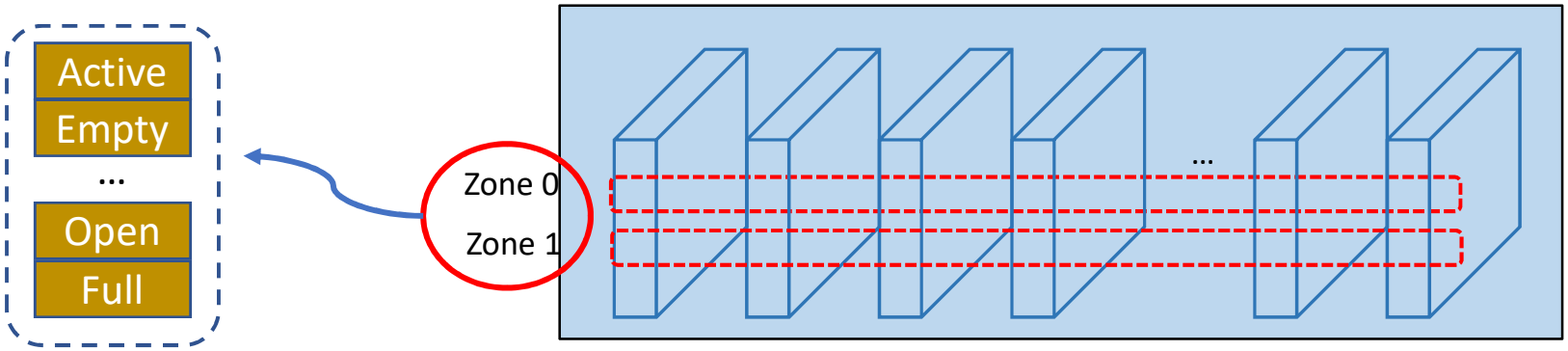




[TLC-based ZNS]

# Challenge 3 : Initialization Time



- **Shortening the init. time of the PB scale system will be competitive.**
  - 128TB * 24ea = 3PB
  - Bring-up time : Server H/W init. + Device map load + PetaOS initial.




**Use zone characteristics as meta info.**

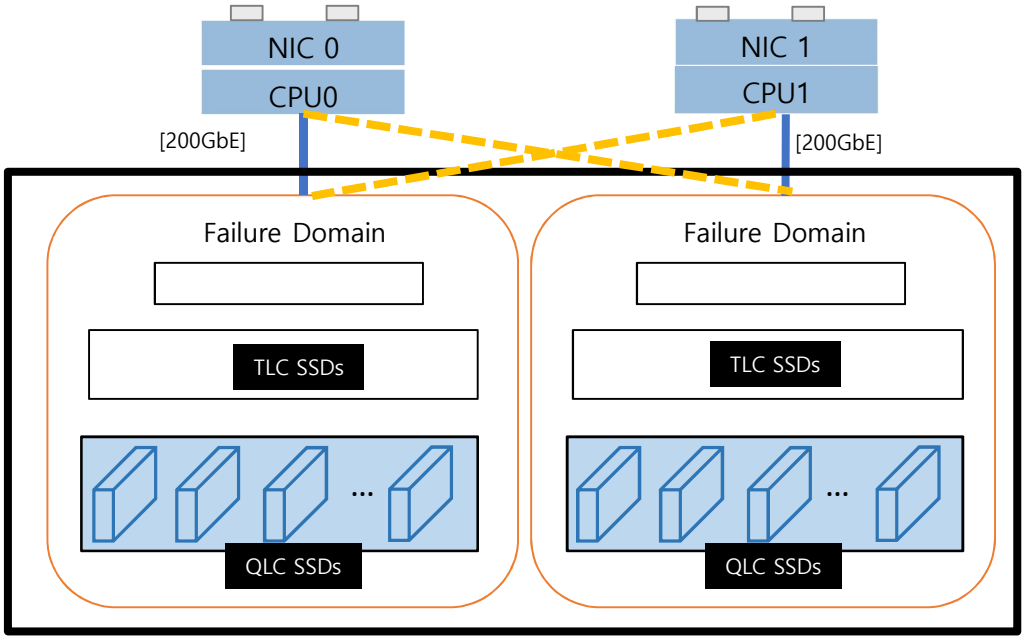
**- Save state info. in the CNS => 4sec**

**- Reconfigure last zone states using limited devices. => 40sec**

# Other Challenges



- **Cheaper Architecture vs. High Performance**
  - CPU sockets, DRAM, DPU, …
- **High-Availability**
  - HW & S/W Co-design
  - In-box vs. In-cluster (multi-nodes)
- **General vs. Customized System**
  - **w/ host file system**



# Summary



- **Questions:**
  - How we can reduce the TCO for NAND-based systems?
  - How the host can reduce the management overhead?
- **Our new storage solution**
  - Increasing storage density to Petabyte scale
  - Offload the host management parts into the box.

SOLUTION

CORE VALUES


**S**peciality

**O**wnership

**L**eadership

**U**pgrowth

**T**ogether

**I**ntegrity

**O**penness

**N**ow

# Thank You